What is the Difference between a Human and a Chimp?

T. M. Murali

murali@cs.vt.edu
http://bioinformatics.cs.vt.edu/~murali

Graduate Research Seminar
Department of Computer Science
Virginia Tech
September 25, 2009
The Story Began in 1953
The Human Genome

- DNA is a (very long) string containing letters A, T, C, and G.
- Length of human genome is 3 billion base pairs.
The Human Genome

- DNA is a (very long) string containing letters A, T, C, and G.
- Length of human genome is 3 billion base pairs (1.8 metres).
- The Human Genome Project determined the spelling of the genome.
The Human Genome

- DNA is a (very long) string containing letters A, T, C, and G.
- Length of human genome is 3 billion base pairs (1.8 metres).
- The Human Genome Project determined the spelling of the genome.
- Eric Lander (Nano-Lecture, 2003 Ig Nobel Prize Ceremony):
  
  *Genome. Bought the book, hard to read.*
The Human Genome Project

Before: human genome has about 100,000 genes.
After: human genome has about 30,000 genes.
The Human Genome Project

Before: human genome has about 100,000 genes.
The Human Genome Project

Before: human genome has about 100,000 genes.

After: human genome has about 30,000 genes.
Shock and Dismay

The New York Times: **Genome Analysis Shows Humans Survive on Low Number of Genes** The two teams report that there are far fewer human genes than thought—probably a mere 30,000 or so—only a third more than those found in the roundworm. . . . The impact on human pride is another matter.

Washington Post: It also raises new and difficult questions, such as how human beings—with all their passions and fears, their capacity for art, music, culture and war—can be all that they are with just 30,000 or so genes, only five times as many as in baker’s yeast.
## Genome size comparison

<table>
<thead>
<tr>
<th>Species</th>
<th>Chromosomes</th>
<th>Genes</th>
<th>Base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>46 (23 pairs)</td>
<td>28-35,000</td>
<td>3.1 billion</td>
</tr>
<tr>
<td>Mouse</td>
<td>40</td>
<td>22.5-30,000</td>
<td>2.7 billion</td>
</tr>
<tr>
<td>Puffer fish</td>
<td>44</td>
<td>31,000</td>
<td>365 million</td>
</tr>
<tr>
<td>Malaria mosquito</td>
<td>6</td>
<td>14,000</td>
<td>289 million</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>8</td>
<td>14,000</td>
<td>137 million</td>
</tr>
<tr>
<td>Roundworm</td>
<td>12</td>
<td>19,000</td>
<td>97 million</td>
</tr>
<tr>
<td>Bacterium*</td>
<td>1</td>
<td>5,000</td>
<td>4.1 million</td>
</tr>
</tbody>
</table>

*Bacterial chromosomes are chromonemes, not true chromosomes

**John Blanchard / The Chronicle**
What Makes a Human Different from a Chimp?

Chimp and chimp genomes are only about 1.2% different!
What Makes a Human Different from a Chimp?

Chimp and chump genomes are only about 1.2% different!
What Makes a Human Different from a Chimp?

Chimp and chump genomes are only about 1.2% different!
What Makes a Human Different from a Chimp?

Chimp and chump genomes are only about 1.2% different!
Genomes provide the parts lists (e.g., genes and proteins) but do not directly tell us how these parts fit.
Molecular Biology → Systems Biology

- Genomes provide the parts lists (e.g., genes and proteins) but do not directly tell us how these parts fit.
- We need to understand how genes, proteins, and other molecules interact with other in different cell states, different tissues, and under different external conditions.
- Study only of individual elements is unlikely to reveal higher-order organisation of cellular interaction networks.
Sea Urchin (Strongylocentrotus purpuratus)

- Very important in developmental biology.
- Many principles of embryo development were discovered in the sea urchin.
Sea Urchin (Strongylocentrotus purpuratus)
Sea Urchin (Strongylocentrotus purpuratus)

- Very important in developmental biology.
- Many principles of embryo development were discovered in the sea urchin.
A Cell
A Cell is a Modular
A Cell is a Modular
A Cell is a Modular Network

Colored boxes indicate post gastrulal domains of expression genes

Endomeso up to 20-24 hours

Veg1 endoderm

Post gastrular terminal or peripheral downstream genes

Late Wnt8 signal from veg2
A Cell is a Modular Network

C Module A functions:

Vegetal plate expression in early development:

Synergism with modules B and G enhancing endoderm expression in later development:

Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with LiCl treatment:
A Cell is a Modular Network that Computes

B
if \((F = 1 \text{ or } E = 1 \text{ or } CD = 1)\) and \((Z = 1)\)
\[\alpha = 1\]
else \[\alpha = 0\]

if \((P = 1 \text{ and } CG_1 = 1)\)
\[\beta = 2\]
else \[\beta = 0\]

if \((CG_2 = 1 \text{ and } CG_3 = 1 \text{ and } CG_4 = 1)\)
\[\gamma = 2\]
else \[\gamma = 1\]

\[\delta(t) = B(t) + G(t)\]
\[\epsilon(t) = \beta \cdot \delta(t)\]

if \((\epsilon(t) = 0)\)
\[\xi(t) = Otx(t)\]
else \[\xi(t) = \epsilon(t)\]

if \((\alpha = 1)\)
\[\eta(t) = 0\]
else \[\eta(t) = \xi(t)\]
\[\Theta(t) = \gamma \cdot \eta(t)\]

Repression functions of modules F, E, and DC mediated by Z site
Both P and CG_1 needed for synergistic link with module B
Final step up of system output
Positive input from modules B and G
Synergistic amplification of module B output by CG_1-P subsystem
Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
Repression function inoperative in endoderm but blocks activity elsewhere
Final output communicated to BTA
**Network is Complex**

The figure illustrates a complex network with various connections and labels. The network appears to be a visual representation of interactions or relationships between different entities, likely in the context of human vs. chimp genomics or related fields. The network is labeled in various segments, with each segment containing detailed diagrams and annotations.

The network is complex and poorly understood, indicating a need for further research and analysis. The figure is a part of a larger project focused on genomics, network analysis, function prediction, infectious diseases, and other projects related to computer science.
Network is Complex but Very Poorly Understood
Challenges with Molecular Interaction Networks

- Biological data sets and networks are large.
- They are intricate and of very diverse types.
- They are noisy: experiments are error-prone.
- They are highly incomplete. We barely know which genes interact, let alone the detailed kinetics of each interaction.
My Research: Understanding Interaction Networks

- Automatically find modules of coherently acting molecules from data.
- Build predictive module-level models of the cell.
- Emphasise a phenomenological approach to systems biology.
- Develop techniques in graph and discrete algorithms, data mining, and machine learning and apply them to solve specific biological questions.
Research Applications

- Predict the functions that genes and proteins perform in the cell.
- Develop drugs that may be effective against multiple pathogens.
- Build models of how cells communicate with each other.
- Zero in on the molecules and interactions that are active in cancer.
- Develop biologically-relevant representations of molecular interaction networks.
Functions of Many Genes are Unknown

- We have the sequences of 100s of genomes.
- We know the locations of genes in these genomes.
- Functions of over 50% of the genes are unknown!
Functions of Many Genes are Unknown

- We have the sequences of 100s of genomes.
- We know the locations of genes in these genomes.
- Functions of over 50% of the genes are unknown!
- Genes with similar sequences in different organisms are likely to have the same function.
- We need techniques for function prediction that go beyond sequence similarity.
Exploit network structure to determine whether gray genes have the same function as the red genes or the blue genes.

Maximally-Consistent Assignments

▶ An edge is *consistent* if it is incident on nodes with the same state.

▶ *Maximally-consistent assignment*: number of consistent edges is maximised.
Maximally-Consistent Assignments

▶ An edge is *consistent* if it is incident on nodes with the same state.

▶ *Maximally-consistent assignment*: number of consistent edges is maximised.

Computational goal: Assign state of $-1$ or $+1$ to nodes with initial state $0$ to achieve maximal consistency by maximising

$$\sum_{(u,v) \text{ is an edge}} w_{uv}s_us_v$$
An edge is \textit{consistent} if it is incident on nodes with the same state.

\textit{Maximally-consistent assignment}: number of consistent edges is maximised.

\begin{align*}
\sum_{(u,v) \text{ is an edge}} W_{uv} s_u s_v
\end{align*}

We used graph cuts to solve the problem.
Human Proteins Interacting with Pathogens

Pathogens are Becoming Drug-Resistant

Pssst! Hey kid! Wanna be a Superbug...?
Stick some of this into your genome...
Even penicillin won't be able to harm you...!

It was on a short-cut through the hospital kitchens that Albert was first approached by a member of the Antibiotic Resistance.
One-Bug-One-Drug $\Rightarrow$ Many-Bugs-One-Drug

- Develop drugs that target *human* proteins.
- Prioritize human proteins interacting with *multiple* pathogens.
Viral Dependency Factors

- RNA viruses like HIV have very few genes.
- Viral dependency factor (VDF): *human* protein that virus needs to replicate and propagate.
RNA viruses like HIV have very few genes.

Viral dependency factor (VDF): human protein that virus needs to replicate and propagate.

- Recent genome-wide experiments have discovered dozens of VDFs for HIV, flu virus, West Nile virus, and Hepatitis C virus.
- Different experiments for HIV show very little overlap.
RNA viruses like HIV have very few genes.

Viral dependency factor (VDF): human protein that virus needs to replicate and propagate.

Recent genome-wide experiments have discovered dozens of VDFs for HIV, flu virus, West Nile virus, and Hepatitis C virus.

Different experiments for HIV show very little overlap.

Proteins in one experiment interact with proteins detected in other experiments.
Predicting VDFs

Collaboration with Michael Katze (Dept of Microbiology, Univ. of Washington) and Brett Tyler (VBI)

- Treat the problem as one of predicting gene function: which human genes have the function of “being used by viruses to propagate?”
- Modify previous approach: for every node $v$ compute a function $0 \leq f(v) \leq 1$ to minimise

$$\sum_{(u,v)} w_{uv} (f(u) - f(v))^2 + \lambda \sum_v f^2(v)$$
Predicting VDFs

Collaboration with Michael Katze (Dept of Microbiology, Univ. of Washington) and Brett Tyler (VBI)

- Treat the problem as one of predicting gene function: which human genes have the function of “being used by viruses to propagate?”
- Modify previous approach: for every node \( v \) compute a function \( 0 \leq f(v) \leq 1 \) to minimise

\[
\sum_{(u, v)} w_{uv} (f(u) - f(v))^2 + \lambda \sum_v f^2(v)
\]

- Solve linear system of equations:

\[
f(v) = \frac{\sum_u w_{uv} f(u)}{\lambda + \sum_u w_{uv}}
\]
Cross-Validation Results

![Graph showing precision vs recall for different methods: SinkSource (BK), SinkSource (B), SinkSource (K).]
Independent Evaluation of Predictions

- VDFs for HIV were discovered in cell lines.
- Protein-protein interactions discovered in a wide variety of experiments.
Independent Evaluation of Predictions

- VDFs for HIV were discovered in cell lines.
- Protein-protein interactions discovered in a wide variety of experiments.
- Used gene expression data from SIV-infected non-human primates.
  - African Green Monkeys (AGMs) are natural hosts for SIV.
  - Pig-tailed Macaques (PMs) are susceptible to SIV.
- Computed which genes were differentially expressed between AGMs and PMs.
- Computed the statistical significance of the overlap between differentially-expressed genes and known/predicted HDFs.
Results of the Evaluation

Blood

Colon

Lymph node

Day 10 Fraction of HDFs

Day 45 Fraction of HDFs
Cells Communicate with Each Other

- Organs are made up of multiple types of cells.
- Proper communication between cell types is essential for optimal organ function.
- Studying such communications in vivo is extremely difficult.
Cell Communication in Engineered Livers

Collaboration with Padma Rajagopalan (Dept. of Chemical Engineering)

- 3D liver mimics composed of hepatocytes and endothelial cells separated by a biocompatible layer of polyelectrolytes.
- Cells appear to communicate across the polyelectrolyte layer, thereby improving hepatocyte function.
Cell Communication in Engineered Livers

Collaboration with Padma Rajagopalan (Dept. of Chemical Engineering)

- 3D liver mimics composed of hepatocytes and endothelial cells separated by a biocompatible layer of polyelectrolytes.
- Cells appear to communicate across the polyelectrolyte layer, thereby improving hepatocyte function.

1. Collect time courses of gene expression in each cell type.
2. Compute response networks in each cell type.
3. Find “hidden” connections between response networks to suggest experiments on specific proteins.
Other Projects

- Building blocks of molecular interaction networks.
- Models to enable explicit comparisons between cell’s response to different conditions.
What Makes Systems Biology Exciting?

- Use principles in computer science to solve problems that impact human life.
- Research is inter-disciplinary: collaborate closely with biochemists, bioengineers, geneticists, doctors, and plant biologists.
- Seek synergistic collaborations between computer science and biology.
- Train students in both computer science and biology.

<table>
<thead>
<tr>
<th>Name</th>
<th>Project/Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Katze</td>
<td>RNA viruses (HIV, flu, etc.)</td>
</tr>
<tr>
<td>Padma Rajagopalan</td>
<td>Liver tissue engineering</td>
</tr>
<tr>
<td>Bruno Sobral</td>
<td>Host-pathogen interactions</td>
</tr>
<tr>
<td>Brett Tyler</td>
<td>Predicting gene function, plant pathogens</td>
</tr>
</tbody>
</table>
How Can You Contribute?

▶ Curiosity about how the cell works, how it becomes ill, how does it survive attacks.
▶ Curious to learn how computer science helps to discover new things about the cell.
▶ We need students who will learn both computer science and biology.
▶ They will form the next generation of scientists in systems biology.

T. M. Murali
Department of Computer Science

Human vs. Chimp
How Can You Contribute?

- Curiosity about how the cell works, how it becomes ill, how does it survive attacks.
- Curious to learn how computer science helps to discover new things about the cell.
How Can You Contribute?

- Curiosity about how the cell works, how it becomes ill, how does it survive attacks.
- Curious to learn how computer science helps to discover new things about the cell.
- We need students who will learn both computer science and biology.
- They will form the next generation of scientists in systems biology.
Nothing in Biology Makes Sense Except in the Light of Evolution